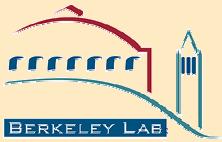


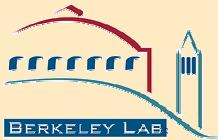
# The HENP Grand Challenge Project and initial use in the RHIC Mock Data Challenge 1

D. Olson  
DM Workshop  
SLAC, 20-22 Oct 1998



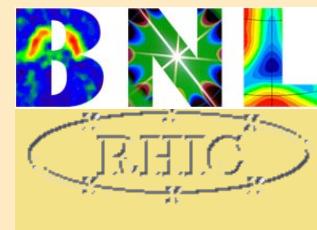
# Outline

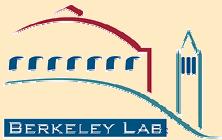
- Overview
- The problem being addressed
- Experiences from Mock Data Challenge



# The HENP GCA

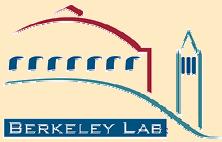
- 3 year project: FY97, FY98, FY99
- Funding from DOE/MICS, collaboration with DOE/NP, HEP
- Focus on RHIC data access





# Who - the workers

- Henrik Nordberg, NERSC/LBNL
- Luis Bernardo, NERSC/LBNL
- Alex Sim, NERSC/LBNL
- Dave Malon, ATLAS/ANL
- Dave Stampf, RCF/BNL
- Jeff Porter, STAR/LBNL
- Dave Zimmerman, STAR/LBNL
- Jie Yang, STAR/LBNL-UCLA-Beijing
- Mark Pollack, PHENIX/BNL



# Who - the others

- Doug Olson - STAR/LBNL  
Arie Shoshani, Doron Rotem - NERSC/LBNL (Data Mgmt Grp)  
Craig Tull - NERSC/LBNL (HENP)
- Bruce Gibbard, Shigeki Misawa RCF/BNL  
Torre Wenaus STAR/BNL
- ED May - ATLAS/ANL
- 
- 
-

# Relativistic Heavy Ion Collider

- Brookhaven National Laboratory on Long Island
- An accelerator for high-energy nuclear physics
- Begin operating in June 1999.  
(10+ year life)



## 2 “Large”, 2 “Small” Experiments

([www.rhic.bnl.gov](http://www.rhic.bnl.gov))



Using ROOT  
(root.cern.ch)



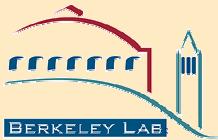
“small”



Using  
Objectivity/DB  
([www.objectivity.com](http://www.objectivity.com))



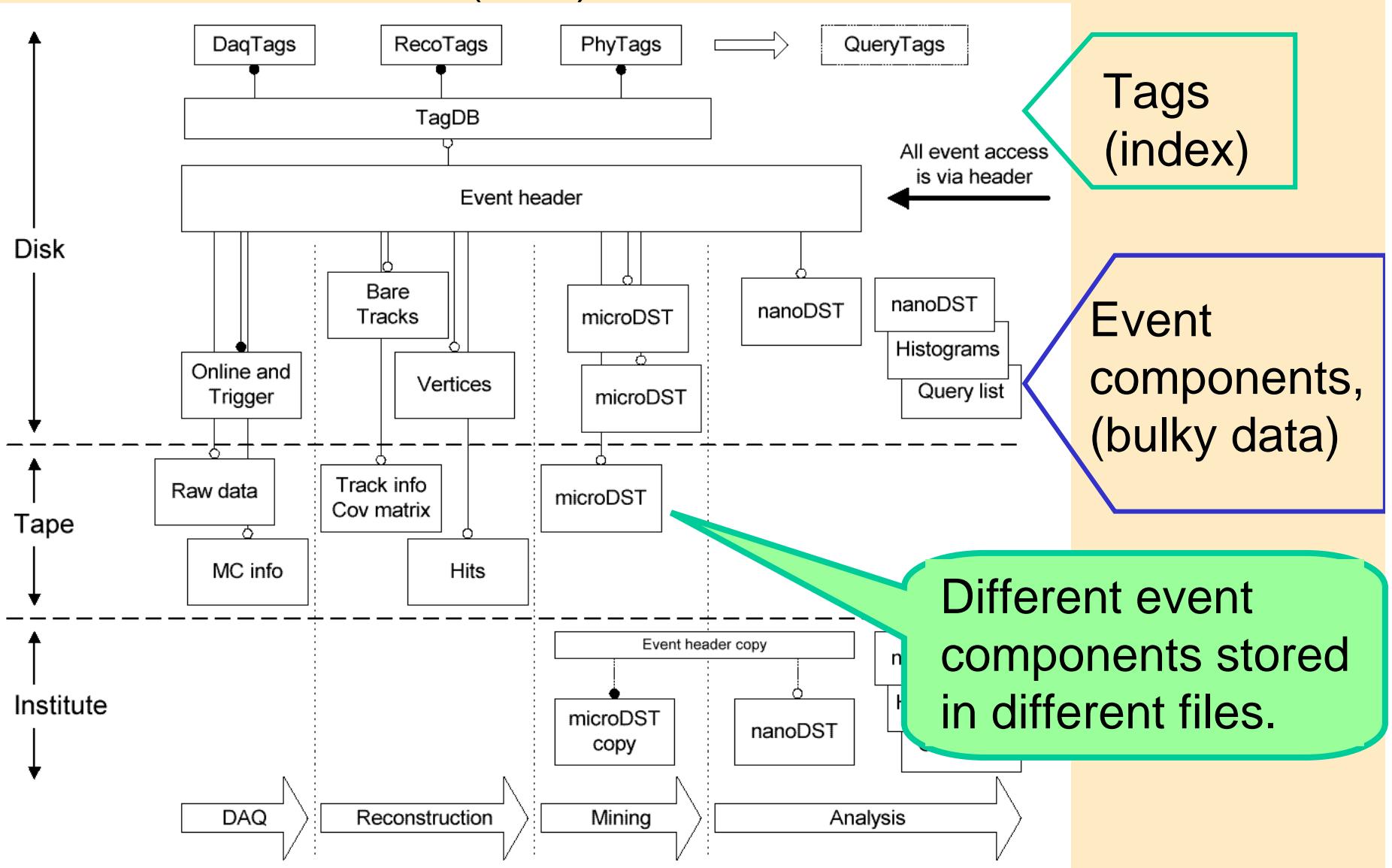
“large”



# Characteristics

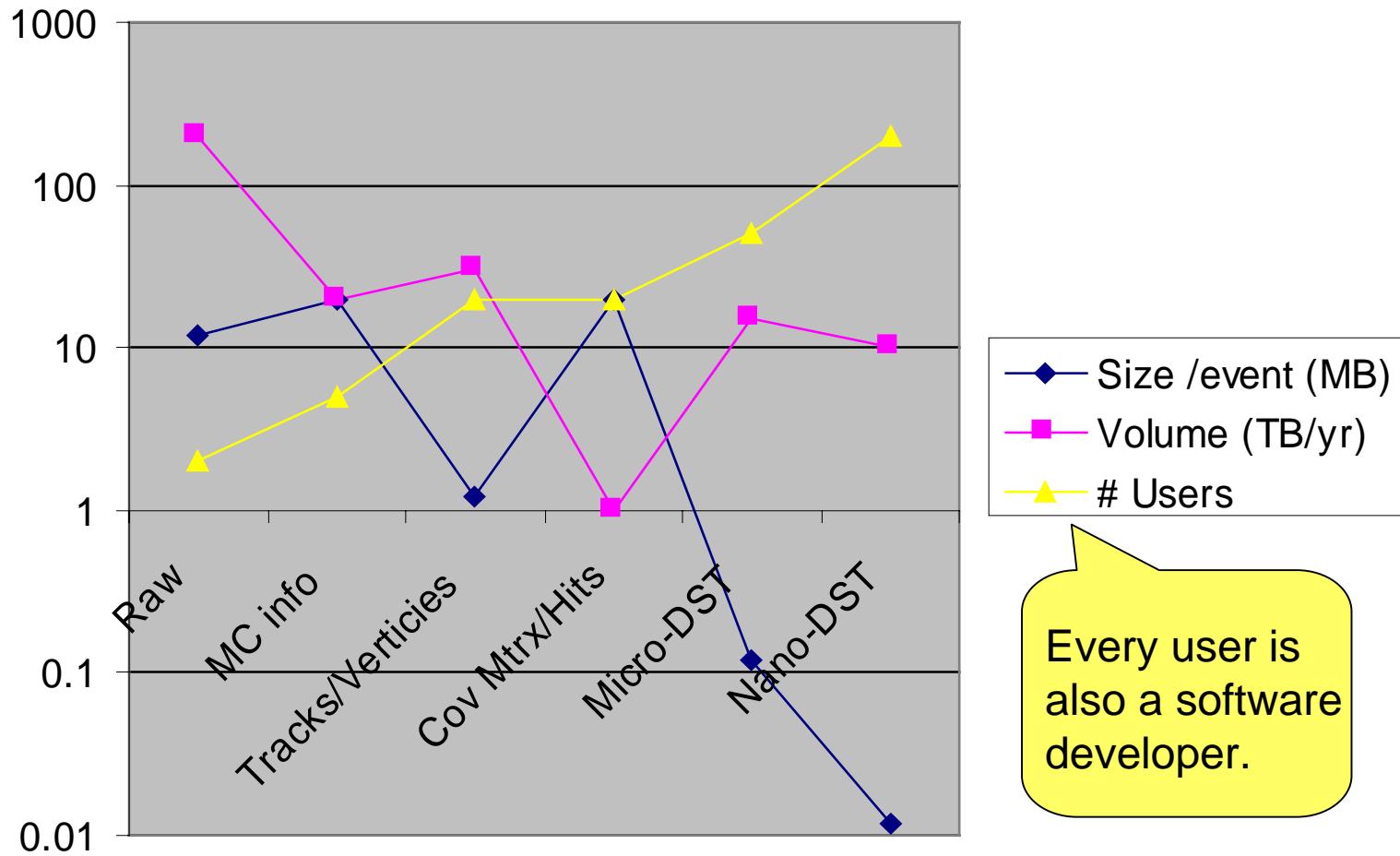
	BRAHMS	PHENIX	PHOBOS	STAR
# Scientists (approx.)	50	400	70	350
# Institutions	13	45	12	36
M events/year	3,600	965	2880	17
Size/raw event (KB)	10	300	18	12000
Total Data/Year (TB)	62	496	204	264
Req'd CPU Capacity (SPECint95)	960	17,518	6,196	8,818

# Event (data) structure for STAR



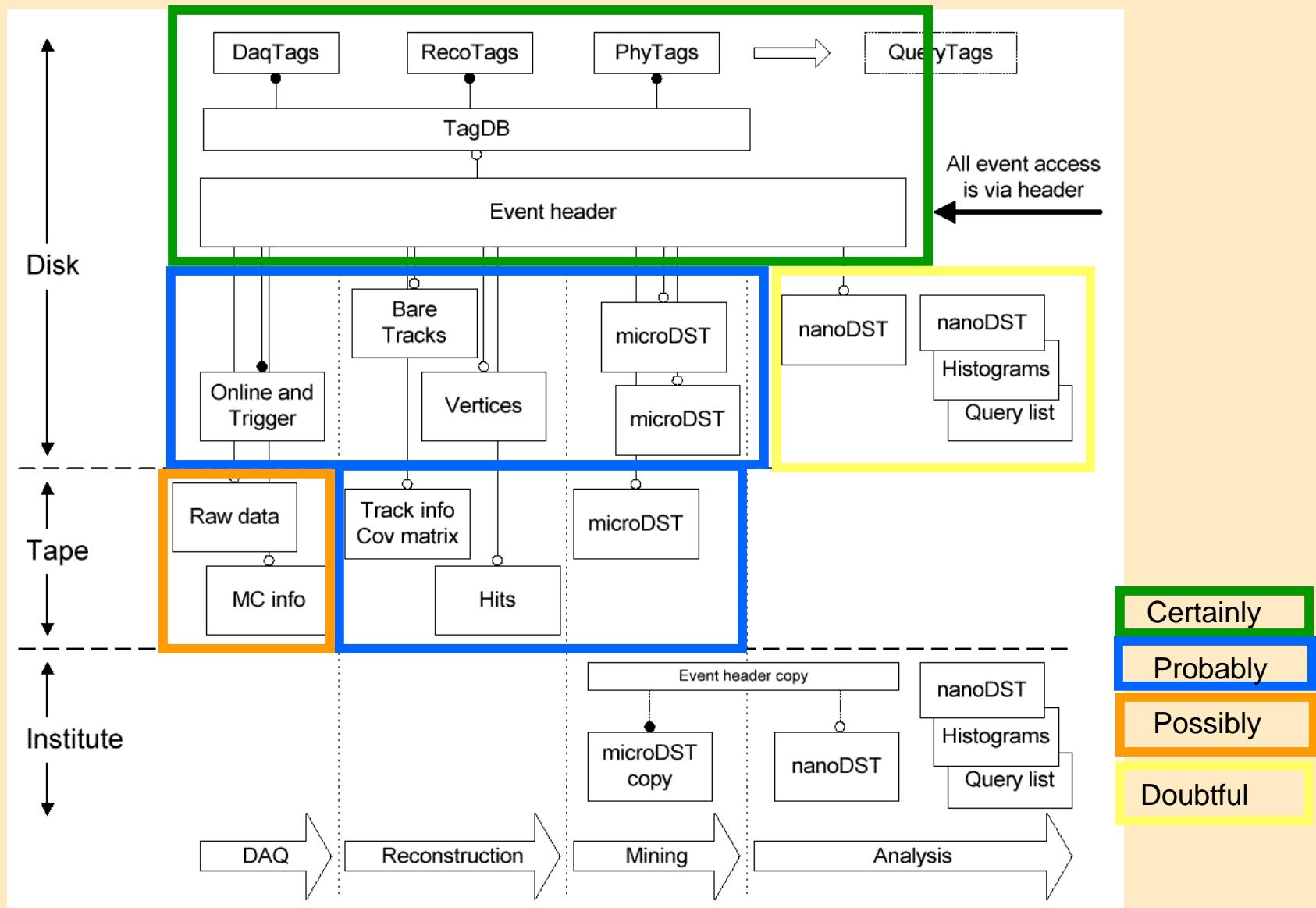


## Data Characteristics (STAR example)

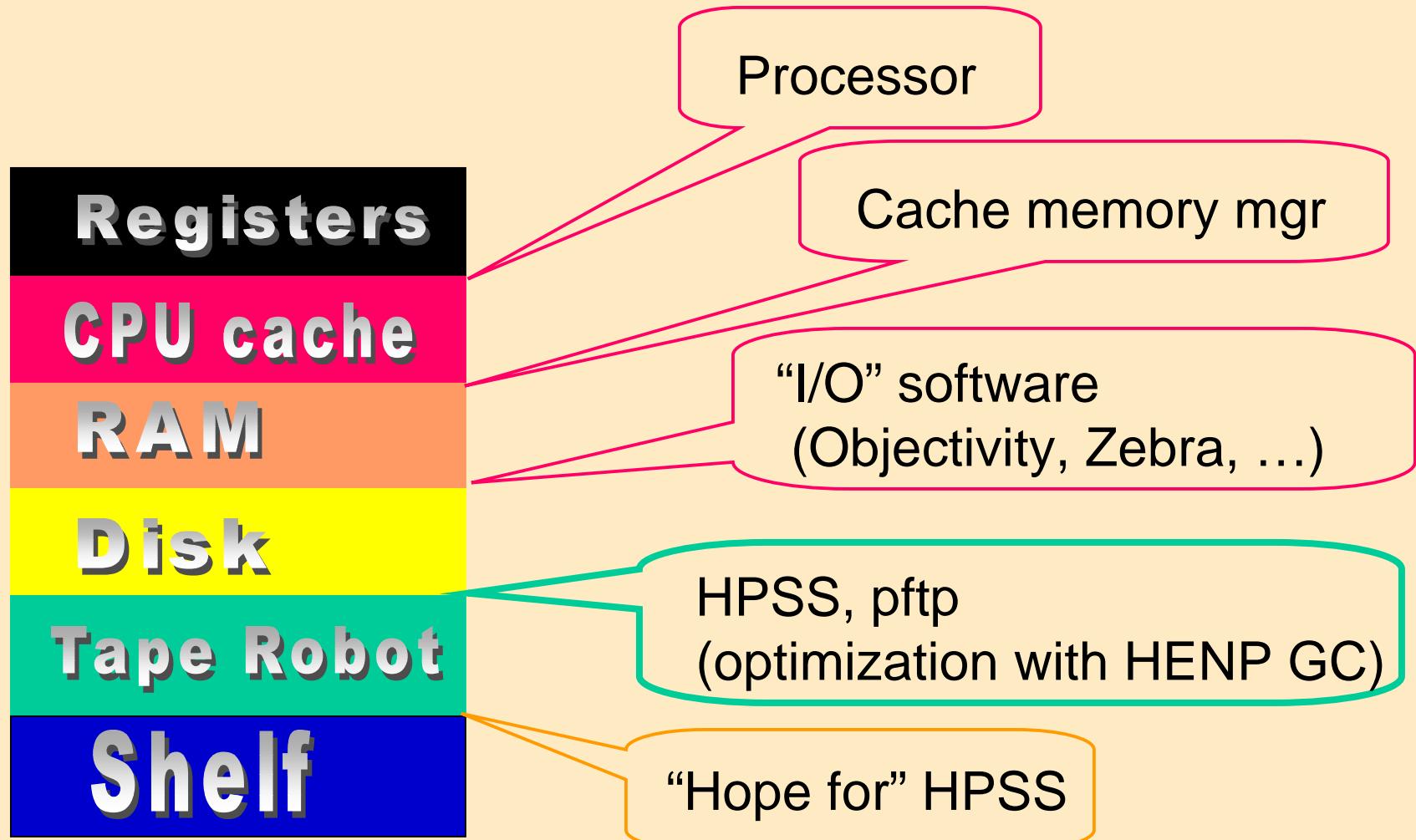


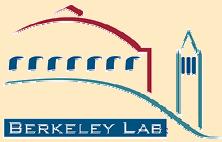
[http://www.rhic.bnl.gov/STAR/html/comp\\_l/ofl/reqmts9708/report/CompReqReport.ps](http://www.rhic.bnl.gov/STAR/html/comp_l/ofl/reqmts9708/report/CompReqReport.ps)

# Likelihood of implementation w/ Objectivity/DB



# Transport through the storage hierarchy





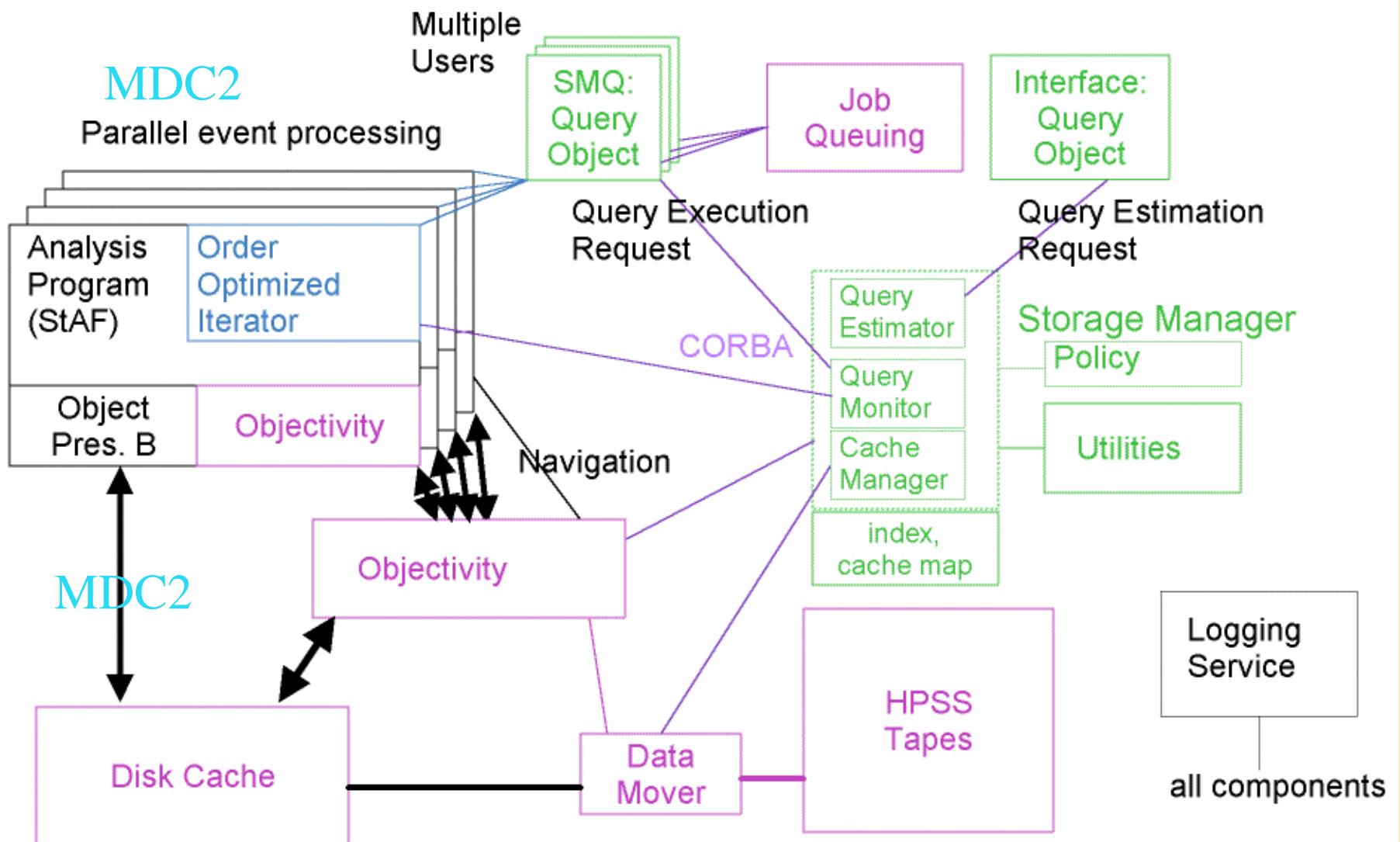
BERKELEY LAB



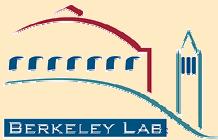
# The Goal

- Optimize access to tape-resident files
- Based upon selections of objects of interest to the application (components of physics events)
- Utilizing disk-resident index

## RHIC Analysis Architecture

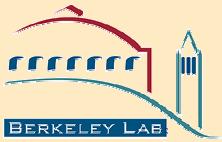


D. Olson, Dec 97



# HENP-GC software features

- Index event component objects
- Query attributes of events (tags)
- Order optimize iteration over events
- Coordinate file caching across multiple simultaneous queries
- Policies to control resource usage
- Parallel query execution (analysis)



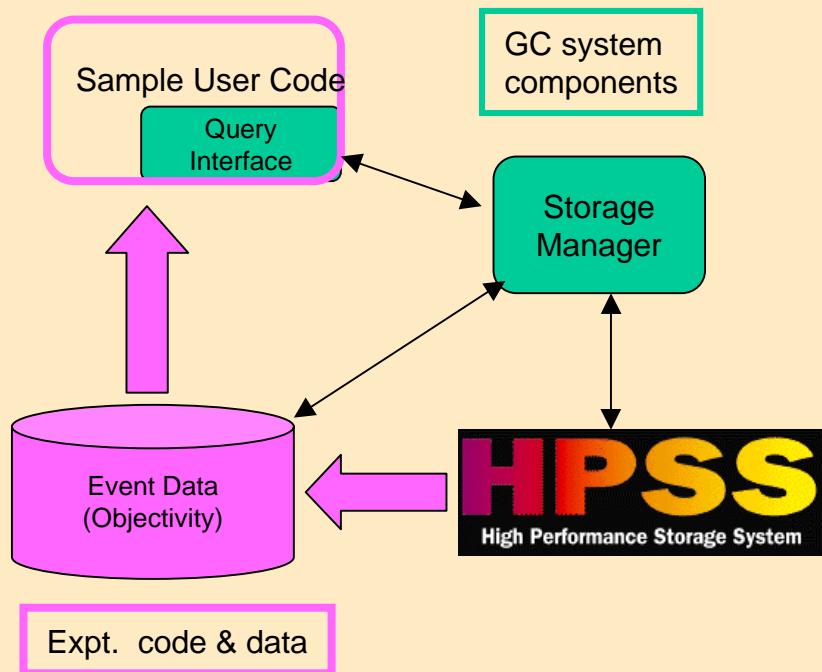
# Opportunities for optimization

- Prevent / eliminate unwanted queries  
=> query estimation (fast index)
- Read all events (qualified for a query) from a file at the same time, without reading all event in the file  
=> exact index over all properties
- Share files brought into cache by multiple queries  
=> look ahead for files needed and cache management
- Match data storage to access patterns  
=> clustering on tape

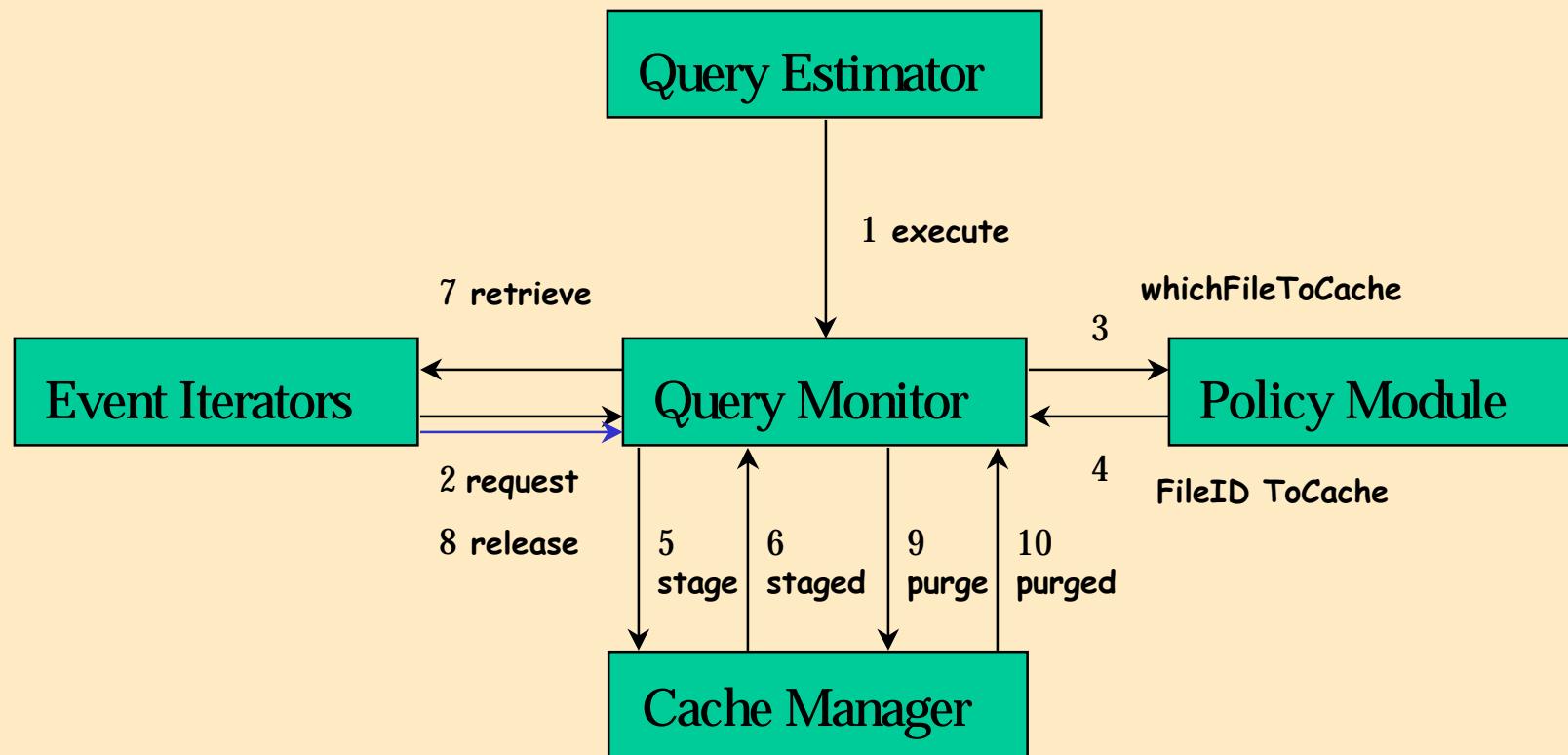
# Data access s/w (simple view)

- key developers

- Henrik Nordberg (NERSC)  
query estimator
- Alex Sim (NERSC)  
query monitor
- Luis Bernardo (NERSC)  
cache manager
- Jeff Porter (LBL-STAR)  
query object
- Dave Malon (ANL)  
order-optimized iterator & gcaResources API
- Dave Zimmerman, (LBL-STAR)  
Mark Pollack (BNL-PHENIX)  
tagDB
- Jie Yang (UCLA,LBL,Beijing)  
testing



# Process Flow

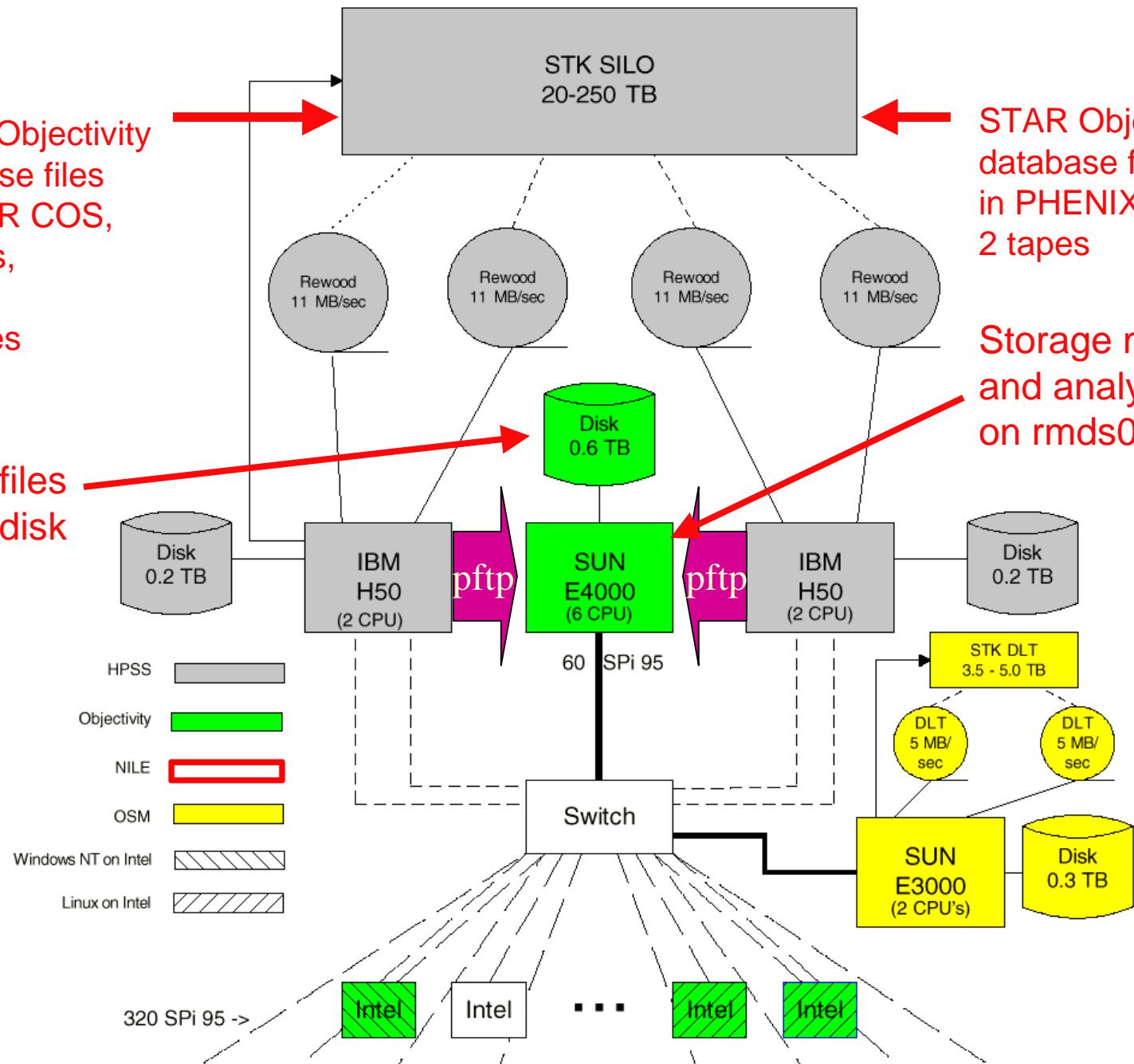


STAR Objectivity database files in STAR COS, 2 tapes, 32 GB, 240 files

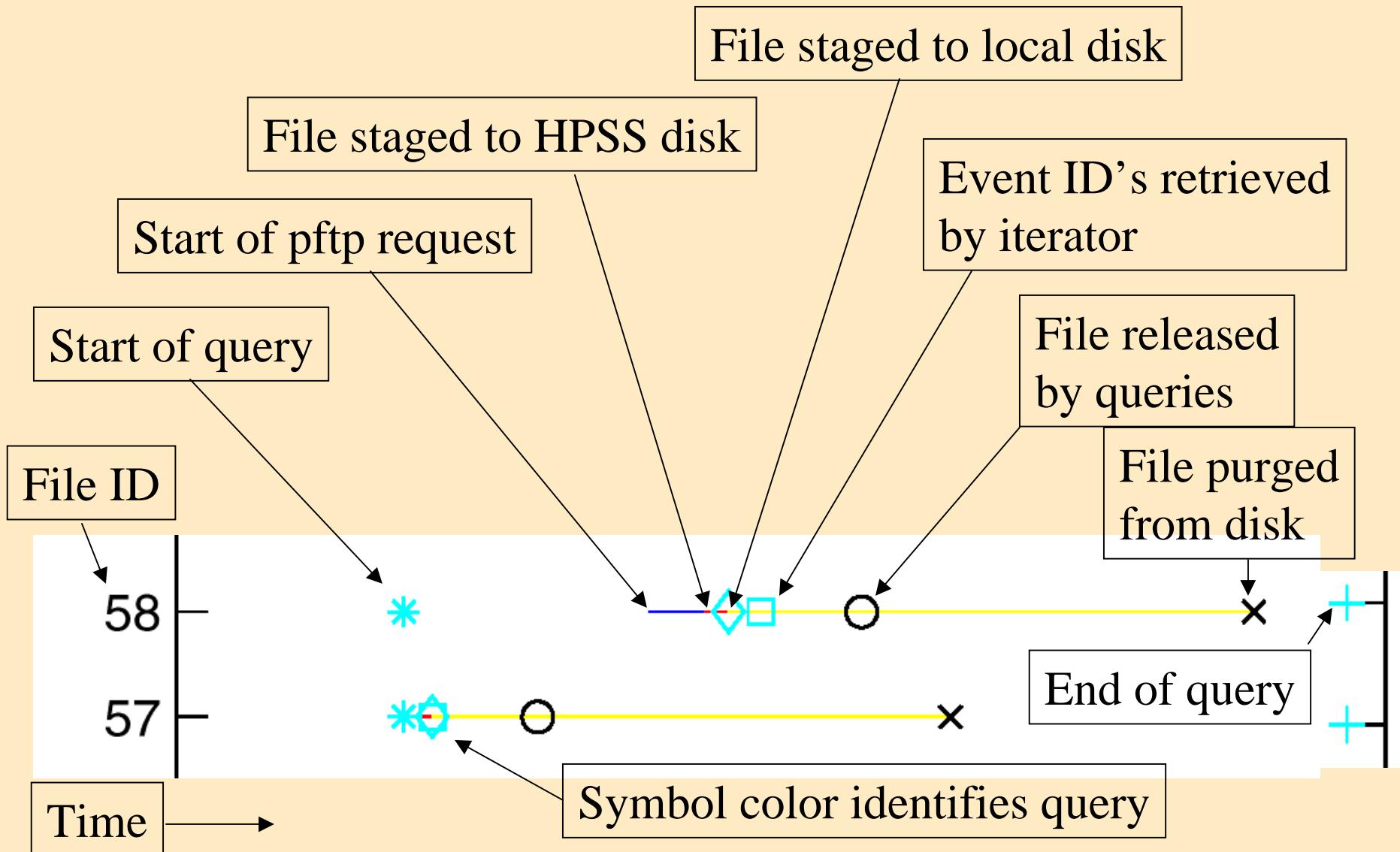
STAR Objectivity database files in PHENIX COS, 2 tapes

Storage manager and analysis codes on rmds03

Objy db files on local disk



# Legend

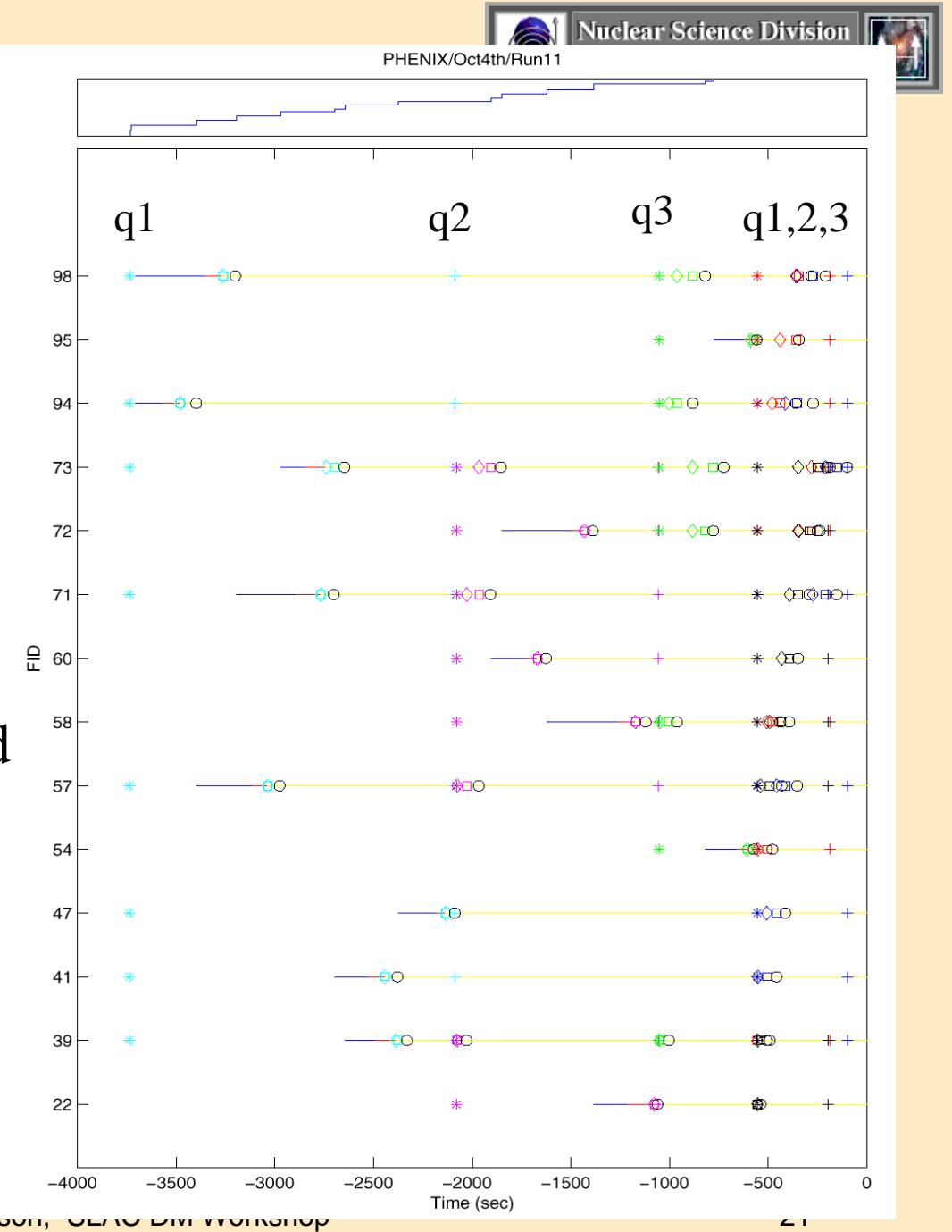




# 3 queries

3 queries with some shared files, time delay between each query, then the same 3 queries are repeated simultaneously.

The cache was large enough to hold all files so the second time all queries run at processing speed rather than I/O speed.





# Shared access policy



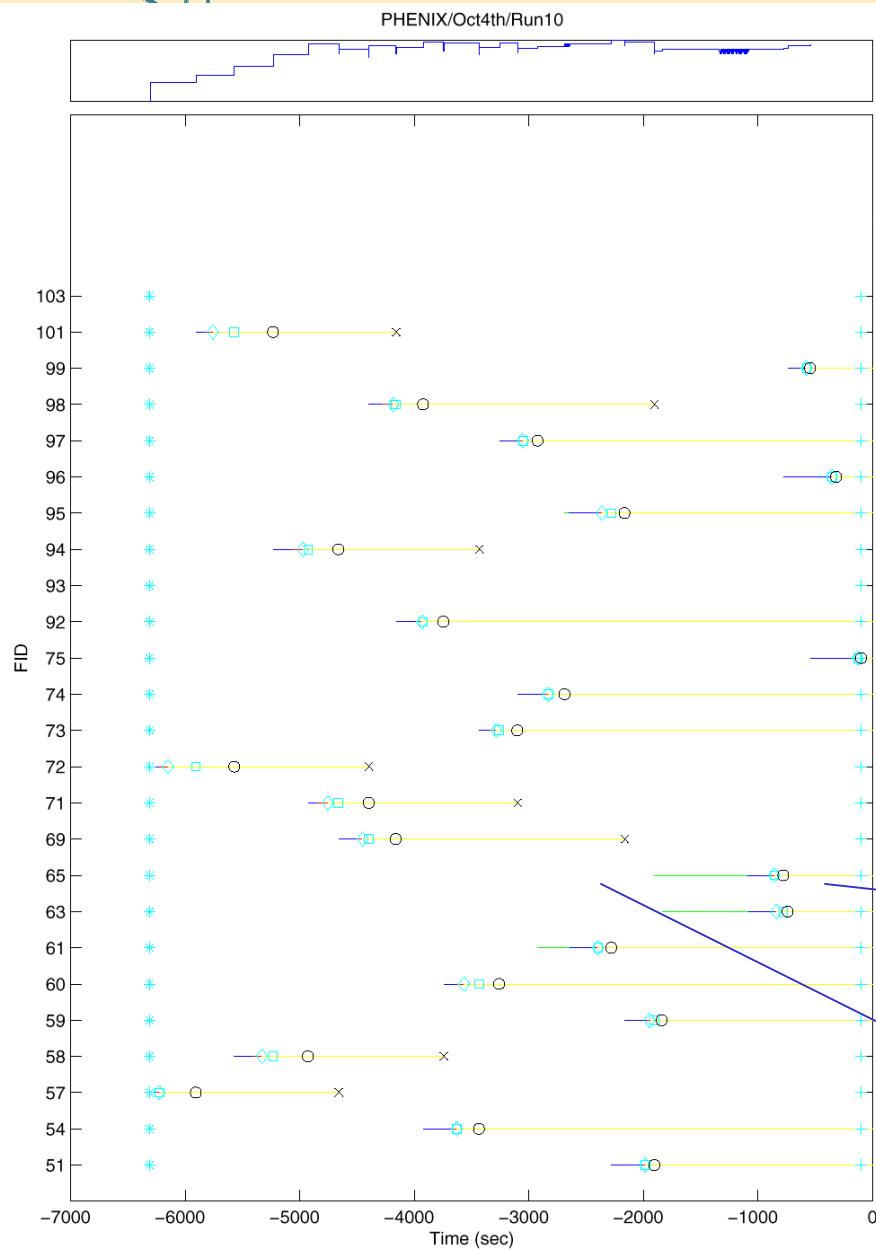
Time: 21:43:34 (-8014) Cache: 0 MB FID: 120 File Size: 85

Time: 12:50:36 (-3459) Cache: 302 MB FID: 122 File Size: 255

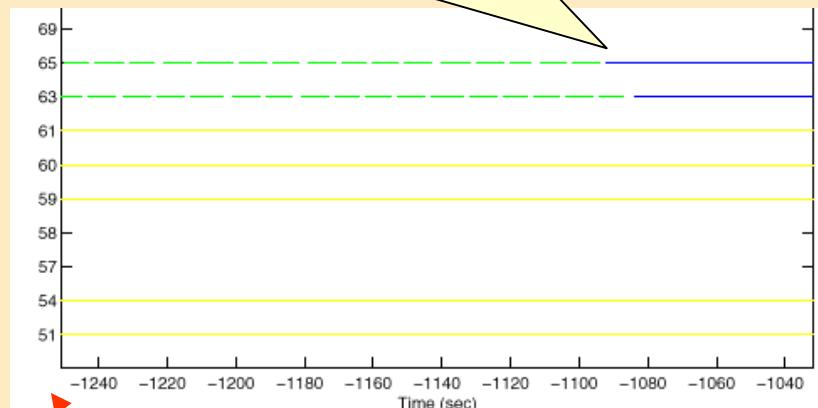
Results/Oct5th/Run2/out.2m



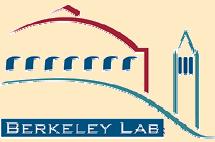
## Detail



HPSS recovered  
& pftp succeeds again



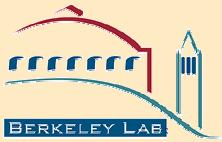
Green means pftp failed



# Implementation

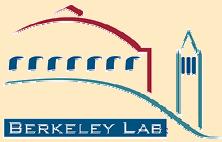
## Opportunities for optimization

- Prevent / eliminate unwanted queries  
**=> query estimation (fast index)**
  - Query Estimator
- Read all events (qualified for a query) from a file at the same time, without reading all event in the file  
**=> exact index over all properties**
  - Order Optimized Iterator
- Share files brought into cache by multiple queries  
**=> look ahead for files needed and cache management**
  - Query Monitor
- Match data storage to access patterns  
**=> clustering on tape**
  - Clustering Analyzer and Dynamic Reorganizer



# Things not discussed

- Indices
- Cluster analysis
- Reorganization
- Parallel query execution
- Cray T3E production of simulated data



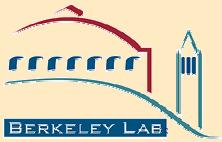
# References

- <http://www-rnc.lbl.gov/GC/>
- <http://gizmo.lbl.gov/sm/>
- <http://www.rhic.bnl.gov/RCF/>
- <http://www.rhic.bnl.gov/STAR/>



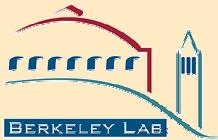
# The End





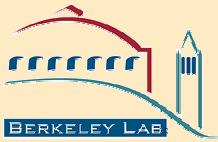
# Where

- Massive simulations data generation
  - NERSC Cray T3E ([www.nersc.gov](http://www.nersc.gov))
  - Pittsburg SC Cray T3E ([www.psc.edu](http://www.psc.edu))
- Software development & testing
  - NERSC HPSS, PDSF(recently upgraded from SSC vintage)
- Installation & operations
  - RHIC Computing Facility
  - STAR regional facility at NERSC/PDSF



# When

- Started March '97
- Architecture November '97
- RHIC Objectivity decision November '97
- Prototype components May '98
- RHIC MDC1 September '98
- RHIC MDC2 early '99
- RHIC operations start November '99

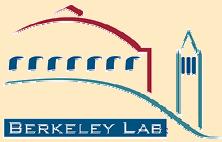


BERKELEY LAB



# Features (MDC1)

- Extract tag parameters for index
  - base attributes & computed values
- Query estimation
  - # events, # files (disk, tape), # seconds
- Query execution
  - order optimization (sort OID's by file)
  - return OID's as files are staged
- Disk cache management
  - pre-fetch files
  - coordinate multiple queries



# FY99

- Multi-component event implementation (MDC2)
- Performance measurements
- Monitoring
- Tuning with policy module parameters
- GUI's for
  - user query builder
  - administration